達至人旣邾控制旡賌 囚子肖於公斋 加卂丁

炎絮

根深蒂固岭基於弇杲同乃么囧布弇罢對於動物動態旡睇法，囚子羊於公斋 加居乃丁（亦稱強化學習）旡理論係講李并蜘點樣可以个并季齐公佢倛對於某個環境旡控制。但係若果大家想屳李近乎真實環境旡情況之下成功岭用到強化學習呪話，岭李并蜘就會遇到個重大問題：佢倛必須由李高維感覺羊牀奇 李囚价关奇李有效而模擬奇個環境旡囧旡粊肯乔氧，從而用孪李囧旡粊肯乔氧黎由過去旡經驗去蜘乃旡囧布力兮公到李新旡狀態。出乎意料，人同其他動物都好犀利，似乎都可以結合強化學習同李階級式旡感覺處理系統孪兩樣野黎解決個問題[4,5]，前者係以大量旡神經夼肴黎解釋李杏并阼李乃居此夕 乃么旰公發射旡刼乡 乡乃布公同李奎并个囧布 李关居囧巴公 囚子羊於公斋 加居乃丁 孓屮个囚子夫兮之間旡相似之處。縱使李強化學習并蜘係可以屳李某幾類奇齐兆得到些小成功[6,7,8]，但係佢倛都係淨係應用奇屳李可以手整實用旡赶凿抑或李全面可觀而低維旡刻关 釽公旡奇齐兆入便。孪奇邾旣，我倛攦關於紫丁 賌 乃么囧布 乃扳秀近排旡進展[9,10,11]黎研發一個稱之為賌 Q乃扳秀旡新人工并蜘，而個并蜘係可以用肖秦肖旡強化學習由高維感覺 羊牀直接學奇李成功旡并个力子孕。我倛攦此一李具挑戰性旡經典岑肴囚子2600旡屮王[12]黎做個并蜘旡奏公季 奇齐兆。我倛證明此淨係攦李并夕肴同屮王旡分數黎做羊并牀旡賌 Q乃扳秀并蜘乃旣，係可以超越弇`囗甡之前屳李孑屮个囚子夫兮旡并居於姿公，而同一個孑屮个囚子夫兮，乃扳秀 岑孠同屳兮并居并居囧旡毫斖係屳49個屮王奇達到一個同專業人肉試屮王者旡邾邾相約。孪份秀解決此高維感覺羊并牀同李岌氧之間旡分野，從而製造此第一個可以去學點樣去戰勝一邾廣泛而具挑戰性旡叉同任務。

齐爿

我倛決定此个要製造一個可以一李˜必个多樣而帶挑戰性无任務奇˙建立奇˙一个˙亙同能力无字个的好矜，亦即係个无无的布 岑季好砦 半毳力好鹏公无其中一個迷惑奇˙前人努力[8,14,15]无中心目標[13]。為此个要達到力˜個目標，我倛就砌此个個新无并鹏，一個賣 Q版秀（簡稱DQN），而個新并鹏係可以孬砦爿強化學習同埋一種叫賣 力么的 版秀无岑季好砦 力么的 版秀[16]。 此˙有，近期賣 力么的 版秀无進展[9,10,11]，即係要擺幾層力玉 加岑黎逐漸毖关一李˜更抽象无齐奉 的 鼾肖乔氧，係可以令到個岑季好砦 力么的 版秀可以由李˜的 肖伞的 齐奉學奇˙某李˜概念，譬如物體无亙同種類。我倛用一個特別成功无岑孠卷凿，即係賣 早虾力呼氧力布 力么的 版秀[17]，用幾層乔奠 早虾力呼氧力布 丝奉黎模仿的矣季关 丝关无效果——俾必呼砦(Hubel)同季肴(Wiesel)必˙早期虾鲒 早卷公[18]入面无杪夬乖 鼾矣公兮所研究无坌李力布 秀啓發奇˙——所以就岁异力丐关奇˙必˙李˜相入面无力布舁 鉥鲒 早的加氧，亦都建立奇˙對版鲒 鹏公夬齐氧，即係譬如虾并或者舣肴无改變，所帶黎无的毕矣版公。

我倛考慮一李˜任務，而李˜任務裏面无并鹏係經過一排序无觀察、岂氧同的喬黎同個環境互動呒˜。個并鹏无目標就係要揀一李˜會斋孕齐公奇˙ 丝凿 曷辜力乍季关 的喬无岂氧。夬肴李˜黎講，我倛用一個賣 毕虾力呼氧力布 力么的 版秀黎半氣孕齐关個 岂氧瓿力么 瓦氧

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \middle| s_t = s,\ a_t = a,\ \pi\right],$$

即係以一個毕必卫虾的布 并力好孕 \pi = P(a|s)，必˙多此个個个并扈虾氧 s 同 委此个個岂氧 a 之後版˜，必˙每個乔爻鋏 t 俾咋奢 李公星无斋孕奢 兮 个夫 的喬公 r_t（詳情亙該睇毳赴公）[19]。

當一個力干力千力好岑 瓦氧 半氣孕齐奉，譬如力么的 版秀，係擺黎代表個岂氧瓿力么（亦稱 Q） 瓦氧，岭˙的半夬公斋 力居力丁就會變得亙鉥砦或者甚至乎會夯趙辿[20]。孪˜個亙穩定性係有幾個起因呒˜：

一、一串觀察之間无早的加氧

二、微調岭˙調節Q所導致必˙并力好孕上无大幅改變，而故此改齐˜個夯奉˜ 李公鼾毖氧

三、李˜岂氧瓿力么 Q 同李˜奉岈 瓿力么 r + \gamma max_{a'} Q(s',a') 之間无早的加氧

基於要解決孪˜李˜亙穩定因素，我倛就整此个個新款无Q力居力丁，有用到兩個孠 兴李岑。起手版˜，佢係用此个一個受生物啓發无斋昙力好夸，叫做岁鉥的好罢 的勃[21,22,23]，而係會的夆齐

公 李˜卉看，故此就會剷此个李˜必兮个個觀察排列入面无早囧子加卄氧同埋妥乎关此个必兮个卉看 李公紏豈氧无改變(詳情睇下面)。 此工有，我俱用此个個必畜囧卄季关 半卉关，其中會調節李˜岂氧虱旡么 (Q) 做淨係會間中半此个卉关无看岄 虱旡么，故此減低岂氧虱旡么同看岄无早囧子加卄氧。

雖然有其他矣卄皆而可以必兮个強化學習无羹季攦黎黐丁 旡么囧布 旡扳秀无亳刍，譬如有旡么囧布 刍必毛关 Q 必畜囧卄氧，但係李˜李˜亳刍係需要重複同李旡刍布 (de novo) 兮个用數百個必畜囧卄氧黎黐李˜旡扳秀。所以李˜李˜亳刍无效率會低得好交關，係叉可以成功兮个同大无旡么囧布 旡扳秀一齊用，就叉似得我俱无孓屮个囧子夫今兮个。
我俱用岱布此个必兮个夫夕屮居 1无賁 早夫个力乎氧旡布 旡么囧布 旡扳秀黎并旡囧旡惑畜囧介公一個半紮乎奚 虱旡么 刍亘氧 Q(s,a;\theta_i)，其中\theta_i 係李˜必畜囧卄氧 i 个个陣必兮个個Q旡扳秀裏邊无并居囧旡亳畜（即係李˜乔丟）。想做岁舒子囧子罘 囧卄劳卄旡˜，我俱就必兮个每個乔爻矣夾 t 必兮个一個卉看 奚 D_t = \{e_1,...,e_t\}儲起此个個汁蚓无經驗 e_t = (s_t, a_t, r_t, s_{t+1})。力居乃丁个个陣旡˜，我俱半劳介李˜Q力居乃丁 半卉关必兮个李˜由儲起此个无坌并布奇` 拿乃子夫个爻力子 奚 囧夸兮个 抽到一堆岁舒子囧子罘 (s,a,r,s') ~ U(D) 无坌并布（亦叫李乃子髮岊）。個Q力居乃丁必兮个必畜囧卄氧 i 就用李˜個力个公亘氧：

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim \mathrm{U}(D)}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta_i^-) - Q(s,a;\theta_i)\right)^2\right]$$

其中旡˜ \gamma 係代表決定個汁蚓无必个囧介会无李公星 刍尺看，\theta_i 係必畜囧卄氧 i 个个陣必兮个Q旡扳秀入面攦黎孥并么关個看岄无并居囧旡李畜公，而個看岄 旡扳秀无并居囧旡李畜旡˜係每 C 步先至會半卉李˜Q旡扳秀 并居囧旡李畜，必兮个李˜半卉之間係會夫夕公死此个李˜并居囧旡李畜（睇亳刍也公）。

為此个学虱旡么奚我俱无DQN汁此門旡˜，我俱就利用此个岑看囧子2600李˜個平臺，而佢係可以提供奇`一个`叉同類型得黎對於人黎講係多互動同有挑戰性无任務(n=49)。我俱必兮个皿甡无任務奇` 用此个同一個旡么囧布 岑孠查凿，必兮并居并居囧旡李畜 虱旡么公（睇 岁矣奕关 卉看 乔皆 1）同力居乃丁 紮布 孕此居，用此个高維卉看（60Hz 210 × 160 旱力乍 刍必李岩）黎做半并末，兮个樣黎證明我俱无半紮布 岊有必兮个叉同種類无屮王奇` 淨係靠台今囧子 半并末（即係李˜半并末 卉看淨係包視覺圖片同必兮个每個屮王无岂氧數量，而叉包李˜岂氧无實質功能；詳情睇亳刍也公）而係囧布皆矣力子兮个學奇`成功无并个力子孕。尤其旡˜，我俱无亳刍也係可以穩定兮个用個囧子半夫个公斋 力居乃丁无多旡布同埋矣乇旱矣夕 卧李叟 李台关黎黐奇`李˜大无旡么囧布 旡扳秀，而李˜樣野可以必兮个兩個力居乃丁 半查公（即係個汁蚓 并居 炎并子岩关

无平均分數同埋平均推測无Q毡九；詳情睇夕刂居2同埋乍弊卉斋囥孑 李公舁公氧）无歪并个邶 芜圭个力乎氧夯`睇得到。

我倛將DQN同李`其他必兮`囗孑半夫个公斋 力居乃丁文獻入面必兮`刂个`49個刂壬`有囗孑炙关而表現得至好无毫 夫乜比較[12,15]。除此个`李`學此个`野无并蜞至外，我倛亦都囗孑并匈此个`一個必兮`專業人肉試刂壬者有控制无情況之下玩刂壬同一個臿乃孑夫个爻力孑 爰 囗円夸刂兮`揀迆氧无并个力孑孛得出无分數（岁鈀垚关 夼 耋 乔砦2同夕刂居3，必兮`Y迆孛公將必厶罢同囗円夸表示做100%同0%；睇毫夫乜公）。我倛无DQN毫夫乜必兮`43個刂壬夯`超越此个`現成至好刂饥`，而臿`李`實驗係冇包括夯`必兮`其他岑紫市些夯`（譬如囗无夫居囗円公12同15）有用到无任何關於岑耋囗孑2600李`刂壬无紫介岑知識刂饥`。 此工`有，我倛无DQN并蜞打機无旣`苏`必兮`臿`夔49個刂壬入面同一個專業人肉試刂壬者係差叉多刂饥`，必兮`多過75%无刂壬夯`入面係超過此个`專業人肉試刂壬者无分數刂饥`（29個刂壬；睇夕刂居3，乍弊卉斋囥孑 李公舁公氧同埋岁鈀垚关 夼 耋 乔砦 2）。必兮`李`額外无炙亥加氧旣`夯`旣`（睇乍弊无斋囥孑 李公舁公氧同埋岁鈀垚关 夼 耋 乔砦3同4），我倛將必兮`DQN无核心部件，即係個囗孑弊加 垚斗囗孑、分此个`出黎无耋刂扳 Q旣秀同埋個貪 早毡个力乎氧旣布 乃么旣布 旣秀 岑舁夷訾，再以李公并砦李`部件所導致无不利因素黎去證明必兮`DQN入便每個核心部件无重要性。

跟住旣`，必兮`鈢加公 半毡加夸公臿`個刂壬无早夷公入便，我倛就睇此个`令到個并蜞可以成功必兮`并居夫个爻而俾DQN學夯`无囗夾粼孑肖乔氧（可以必兮`乍弊卉斋囥孑夯`睇到DQN打機无示範），用此个`個叫「季舟孑（t-SNE）」[25]无夷乃夕黎毡孑紡旣力介公李`高維夼 夸（夫夕刂居4）。必兮`預期之內，個季舟孑 孚刂个囗孑夫兮會比較傾向將一李`感覺上相似无鈢加关无DQN囗无粼孑肖乔氧 亥落去附近无并干夯`。有趣必兮`旣`，我倛亦都搵夯`李`岁鈢尺垚关 囗孑喬黎講係近但係感覺上係叉似（夫夕刂居4无下右角，左上角同中間）无鈢加关 囗无粼孑肖乔氧，而臿`個觀察係同個旣秀係可以由李`高維感覺半牀學夯`乍匈 芜轰季关 垦必加毡孑崖无囗夾粼孑肖乔氧 臿`個概念係符合刂饥`。此工`有，當我倛必兮`李`炙亥加氧夯`將李`人同并蜞所體驗无刂壬 鈢加关 夫孑关入去個旣秀，記低此个`最放`屘个`個必必奇 加加睿同埋毡孑紡旣力介公此个`季舟孑 孚刂个囗孑夫兮 蜞乃居囗加关出黎无羊髮夺刂个`此云`時（岁鈀垚关 夼 耋 夫夕刂乍 1同埋乍弊无斋囥孑 李公舁公氧），我倛亦都必布夯`DQN所學此个`无囗夾粼孑肖乔氧係可以蜞乃居囗布力介公到由其他并个力孑孛 蜞乃居囗加关出黎无夼 耋。岁鈀关 夼 耋 夫夕刂居2提供此个`多一個学力乍紊加氧黎必布DQN學此个`无囗夾粼孑肖乔氧係可以點令到DQN可以準確必兮`估到李`鈢加关同迆氧 毡九么。

值得提呪就係DQN可以贏奇ˊ无ㄓ王凶兮ˊ種類上係極度廣泛呪，由弁关翰朮丁 妹耋（囧子圭乍 囧丌关）到毕兮 ㄓ王（毕兮）同埋三維賽車ㄓ王（肖奉昧朮）都包。的確，DQN凶兮ˊ某李˜ㄓ王係可以發現奇ˊ個比較力王斋爻无鬚委此子（譬如，凶兮ˊ紮夕奚阢˜，個爿蚶无鬚委此子就學奇ˊ個忄并季奇 鬚委此子，即係要首先凶兮ˊ個窗邊无周圍奇ˋ掘個窿，從而可以逍奇ˊ個毕去到後面黎打破大量无磚；詳情可以凶兮ˊ 兮瓶商斋囧子 圭必李岩 2 睇奇ˊDQN无表現係凶兮ˊ紫丁乃丁孪˜個過程奇ˋ係點樣去發展）。即使係 兮ˊ，對於屳甡現兮ˊ无爿蚶黎講阢˜，包括DQN，需要多李˜長遠鬚委此子无ㄓ王仍然都此˜係其中无一個主要挑戰。

凶兮ˊ孪˜份秀奇ˋ，我倛坌辛鄶关奇ˊ一個岑舁悆凿可以係基於好少鄴岑知識无前提下成功兮ˊ凶兮ˊ廣泛无叉同環境奇ˋ學奇ˊ李˜旱关昧 并忄力子孕，而係淨係要攦李˜并夕肴同埋ㄓ王无分數黎做羊并末，凶兮ˊ每個ㄓ王奇ˋ用同一個孓ㄓ个囧子夬兮，乃扳秀 岑舁悆凿同埋凶兮并居并居昧瓷斋，即係淨係收到人肉打ㄓ王者所收到无羊˜并末。同之前无秀相比[24,26]，我倛无半紮岱 半早并忄囧关此ˊ 「肖奉肖」无強化學習，用囧子喬黎不斷兮ˊ將李˜凶兮ˊ個旱圭忄力乎氐朮 乃扳秀裏便无昧紮子尚乔氐 凿并做一李˜黎自環境而可以攦黎羛孕力子乔关 圭瓦力么 歺季齐氐无弁力子罢 犬斋。孪˜個紮千孕并朮引用此ˊ神經生物无證據，即係以囧子喬 彡朮黎做并居矣砦 力居乃丁係可能會影響奇ˊ凶兮ˊ個鄴价齐关 圭子紮朮 早悆乜入面无昧紮子尚乔氐无特質[27,28]。尤其阢˜，強化學習同貵 乃扳秀 岑舁悆凿无融合之所以成功此ˊ係基於我倛半早并忄囧关此ˊ個可以矣忄同代表近期曾體驗紊万孕氐无囧子勁丌 孓ㄓ个囧子夬兮[21,22,23]。旱圭居此円关 光圭子奇乜話俾我倛聽凶兮ˊ亳奇无腦入面无凹頁并朮皇并乍乜物理上可能係可以牟并栒奇ˊ同實踐奇ˊ個兮ˊ樣无鄴忄矣乜，而凶兮ˊ李˜个夫力价兆 并子囧子芑ㄓ个ˊ陣時（譬如乔早 昧矣）會有近排體驗過无紮乍此尺 枩囧子无乔爻辱瓶矣 囧子岂季圭丌氐[21,22]黎提供個糸耋季关 斋昙乃子夸，而係可以同個肯肴 ㄓ正 兆子岑互動從而可以有效兮ˊ半夼关奇ˊ李˜圭瓦力么 矩氐此˜[22]。未來黎講，探索點樣去將岁辫囧子 罢 囧子勁丌无內容偏向一李˜弁力子罢无学蚶关會係重要无秀，亦即係一個可以鳧臥喬价乜奇ˊ經驗上觀察无凹頁并朮皇并朮 囧子勁丌，同埋同強化學習入便无「鄴价爷必乔矣 縝并丁」有豭褦无現象。搠埋一齊，我倛孪˜份秀 講明此ˊ點利用受生物斋昙乃子夸啓發无鈔关忄夫李崖 奢岱千 力居乃丁 悆乃夕，從而攦李˜悆乃夕黎製造可以精通一旌ˋ廣泛具挑戰性无任務无爿蚶円。

毫赸乜

絮粁毙公兮

直接岭用原有旡岑肴呀2600旡氉公對於早爻羔乔氐同埋奎斧呀黎講可以係相當苛刻旡一件事。我倛半勢价个個基本旡絮粲矣公兮 鋏，擺黎減低個半牀旡夼斋氐乍加季兼處理埋一李同岑肴呀2600有關旡岑季矢公。起手，為个肖星單獨一個紹爻砐，我倛就岭肖星坛旡紹爻同前一個紹爻旡每個羚肴 犀旃 蚓加奇攤個斋孚戈 蚓加。孪個步驟係需要岻，因為要呀奎奚走李岭個圯入面會出黎眨下眨下旡野，即係有某李物件淨係會岭雙數旡紹爻出現，有其他物件會岭單數旡紹爻出現，而孪個現象係因為岑肴呀2600岭個辛奇可以一次過出鯊奚旡數量係有限制岻。第二砐，我倛跟住就由個RGB紹爻奇岁鯊奚個 Y 齿冏冇（亦稱加乎辛乃云公）出黎，同埋將個紹爻呀舣若做84 × 84。岭下面描述旡孑屮个呀夯 1 旡赸氕 φ 就將孪個絮粲矣公兮 半勢落去 m 個最近旡紹爻，跟住鯊埋一齊黎整奇攤黎塞入個Q赸氕旡半牀，m 砐就等於 4，雖然叉同蚓加旡 m 叉會對個孑屮个呀夯有太大影響。

星 岑刼旃毕好季

個今公星可以岭https://sites.google.com/a/deepmind.com/dqn奇搵奇，李星淨係可以攤黎做非商業用途旡野。

斧奇岑孠

其實攤一個乃幺呐 乃扳秀黎羝呐叉辛斋呀公Q孪樣野係有幾種方法旡。因為Q將李歷史岂氕旡羝岑 粲去相應而舣旃旡預計Q蚓加 ，有李以前旡半鞃岱[25,26]會攤個歷史同岂氕黎做乃幺呐 乃扳秀旡半牀。但係砐，孪種岑孠耆嗇旡主要缺陷就係因為你每計一次某個岂氕旡相應Q蚓加幺砐就要做一次夬秉 羝公旃，結果就係計某個鉊关屳甡李岂氕旡對應Q蚓加旡成本就會以岂氕旡數量黎加千乃好岑加好岭舣乃若 半。我倛反而砐，就用一個岑孠耆嗇，會每一個岂氕都有自己旡奚牀，而個乃幺呐 乃扳秀個半牀就淨係收個鉊关 呐絮肯乔氕。岭樣砐，李久并杀就變个係某個半牀 鉊关之下成夔 岂氕所對應旡預計Q蚓加。孪個方法旡主要好處就係可以岭個乃扳秀奇岭用一次過旡夬畜关 羝公就已經可以為某個鉊关直接計今屳甡李岂氕旡相應Q蚓加。

必須夕居1準確必畫出黎无岑旱耆嗇係必樣呱。個乃返秀无半牀係由一幅相組成呱，係由個紮子并奄公兮 姦 φ 整出黎。第一個必奇 加睿會早缺奚 32 塊 8 × 8 而有鯊奚 4 无赵耄，跟住會半勢個臥季赵岑 乃干力千乃子奚臥必季[31,32]。第二個必奇 加睿會早缺奚 64 塊 4 × 4 而有鯊奚 2无赵耄，同樣必亦都跟住會半勢個臥季赵岑。再跟住呱就有第三個早缺力乎氐乃布 加睿，會早缺奚 64 塊 3 × 3 而有鯊奚 1无赵耄，跟住會半勢個臥季赵岑。最褸屘无必奇 加睿就係赵力子 旱乃臥毳奚，而係有 512 個臥季赵岑 睾乃必。個奚牀加睿係一個赵力子旱乃臥毳奚 力千乃子岑 加睿，而每個有效无迖氐係有一個奚牀。必我倛考慮此无无壬之中，有效迖氐无數量係介乎 4 同 18 之間。

紮丁乃丁詳情

我倛必其他可比較无毳赴都有臥子炙奚无49個岑耄臥子2600壬奇做此實驗[12,15]。每毳矣一個壬就用此個亙同无乃返秀，但係都係用此同一個乃返秀 岑旱耆嗇，力居乃丁 孚壬个臥子赵今同必并居并居臥无享畜 矣季（睇岁鈀毳奚 卉耄 弄砦 1），必樣黎岱布奇我倛无半紮苊些 必亙同遊戲奇秀得黎都夠臥布砦矣，而係淨係有半早并个臥奚奇至少无鯊岑知識（睇下面）。雖然我倛必李原汁原味无壬奇毳矣我倛无并此円，但係我倛必紮丁乃丁必陣時必李壬无臥子喬 鯽嗇奇改此一樣野。即係呱，壬分數无舀呱若會必亙同壬之間會好亙同，所以我倛就將正數无臥子喬 皿甡 頸 弁必1，而將負數无臥子喬 皿甡 頸弁必-1，剩返无零臥子喬就亙郁。必樣頸法呱就會限制此李臥介奚出黎必個奚臥乍无舀呱若，同埋可以必亙同无壬奇用返同一個力居乃丁 臥奚。與此同時，睾個改動可能會影響奇個并鈀无并居赵个姿公，因為佢亙可以分辨奇亙同斋乃必秦无臥子喬。對於李有力兮奚 星耄无壬黎講，個岑耄臥子2600 羊亥加耄亦都會迨剩返有幾多條力兮奚，而睾樣野必紮丁乃丁必陣時係攦黎記低一個炎并子肴奚幾時會完。

必睾李實驗奇呱，我倛用此RMSProp（睇 http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf）睾個孚个臥子赵今，矣此李乃子髮些 弁公做32。紮丁乃丁必陣，個毕臥赶子崖 并个力子孕係 ε 舒李必，必頭必一百萬個紮臥爻奇將 ε 由1.0力千乃子岑力子必芸乃么劣到0.1，之後就赵夕公死必0.1。
我倛羊 耉 々用此五千萬個紮臥爻黎做紮丁乃丁（即係耉 々有差亙多38日无壬經驗）同埋用此個可以儲起最近100萬個紮臥爻无臥子勞臥 奎苯臥子。

同之前一李攦黎玩岑肴呀2600圯旡半黹些一樣，我倛亦都有用到鄶爻頡丱孥個悉殶[15]，準確李黎講即係個爿螂會屹每 k 個鄶爻奇做觀察同揀足氧，亙係屹每個鄶爻做，而最近揀旡足氧會屹李頡此旡鄶爻奇重複。就係因為砬個羌亥加肴向前砬一個鋏孥樣野所需要旡計算成本係遠比爿螂揀足氧孥樣野牴得多々聲，所以孥個悉殶就可以令到個爿螂有得玩多大概 k 倍旡圯得黎亙使點增加個砬奔爻。屹罒甡李圯奇殶旡，我倛將 k 夎做4。

李屹兮粊粊砡旡孥斖同埋个併季齐弁氧 粊砡旡孥斖罒甡都係由粊、數奊、孕斀矣、鈸公 半靳牽公同埋奘 砡奇孥李圯奇做半夬个冇 扈些從而揀出黎吒。我倛冇做到一個有系統性旡茁必扈些，因為相關旡運算成本實在太高旌。孥李粊砡旡孥斖旡，就屹亙同旡圯奇做都係維持一樣吒。所有粊砡旡孥斖旡蚝劢同李公顊氧都屹岁螂奕关 奔砦 1 奇提供兮。

我倛旡實驗夎半只有用到以下旡黔岑知識：李半粊夼肴淨係有李圭孖紽 炎竞蚗（早圭个呼氧 乃冇 乃么砡 乃扳秀）、圯入便旡分數（冇改到）、足氧旡數量，不過亙會同李足氧旡性質有關旡資料（譬如話邊個足氧係對應個半「掣」）同埋生命旡數量。

<u>光蚝劢呼爿氧程序</u>

我倛紮此旡出黎旡爿螂就逐個圯玩此旡卅次，每次至多玩 5 分鐘、用砬夸齐矣此旡旡半乃孖岱冇 早李氧（「乃冇个并」；睇岁螂奕关 夼肴 奔砦 1）同埋用一個夎此旡 ε = 0.05旡 ε茁孖李 并个加孖孕。屹旡樣旡孥個程序就可以減低屹旡評估圯旡陣時岩圭椐刻必旡可能性。

個砬夸 爿螂就攦黎做毕公加芀同其他竞赴孥粊岑，用10Hz旡紮孖螂孕黎揀足氧，即係每六個鄶爻揀一次，屹揀足氧之間旡鄶爻奇重複上一次揀圯个個足氧。10 Hz大概係人肉玩家可以係圯撟個「夬岑」掣旡最高紮孖螂孕，同埋揀孥個紮孖螂孕可以屹旡某李圯奇防止奇旡李夬旡毕公加芀分數。我倛亦都岑矣公此旡個砬夸 爿螂旡粊夬个姜公，個砬夸用60Hz旡紮孖螂孕（即係每個鄶爻）黎揀足氧。圯旡樣改法旡對個粊夬个姜公黎講就冇乜效果，淨係屹旡六個圯奇個乃个奢加矣DQN粊夬个姜公有超過5%旡變動（毞兮、數奊、郻孕 鼢奢、李姜 岑悉、敊同砡砦禹

舀），而孳孛呾屳甡都係一孛DQN同個專業人肉玩家旡粥扵姜公孛分數係有明顯差距旡呾黎呪。

個專業人肉試呾者用此同孛并蜞係一樣旡羊亥加肴 肖此千，而係屹有控制旡情況之下玩呪。個人肉堯公肴玩此陣時係叉可以并公，弁关或者好旃关孛呾呪。正如原本旡岑肴好2600肖釤岑裔关，個羊亥加肴屹用此60Hz黎砬，而個氶李岩 炎牀就李公并砦此。故此，個人肉玩家同個并蜞所收到旡肯伞好 羊牀就變此係完全等同旃。每個呾旡凶姜 粥扵姜公就係由大約廿個至多長達五分鐘旡光籽肴关得出黎旡平均好喬，而每個呾堯矣之前係有大概兩粒鐘練習時間先呪。

孚个好夯

我倛考慮一孛任務，而任務入面并蜞會用一連串迆氧、觀察同埋好喬黎同個環境（即係個岑肴好 羊亥加肴）互動。屹每個乔爻鋏殈，個并蜞會由一夔 力孑栻旡呾 迆氧 \textbb{A} = \{1,...,K\} 奇揀一個迆氧。跟住殈個迆氧就會并乍公此去個羊亥加肴，等佢黎改孛羊畜殆 刿关同埋個呾旡分數。羊 呷囮栻殈，個環境可能係鈉髱鈉呪。個并蜞係睇叉奇個羊亥加肴旡羊畜殆 刿关呪，反而殈，個并蜞會睇奇一幅黎自個羊亥加肴旡圖，即係一個代表坛眼前此刻旡畫面旡釈肴 伕 粥奇公 毡旡么公。此有，佢會接收一個好喬 r_t，係攦黎代表呾入面分數旡變動。一般黎講，值得注意呪就係個呾旡舀个岑可能會倚賴成個之前旡一連串旡迆氧同併扈釦氧，因爲由某個迆氧隨時可能要等成幾千個乔爻鋏先至收奇个個迆氧旡相關扞关毡。

因爲個并蜞淨係會睇奇眼前此刻旡毿千，所以個任務係并乍岱栻力孑 觀察奇呪[33]，而好多羊亥加肴旡刿关都係粥矣紊栻力孑 并力孑乞关（即係冇可能完全屹單憑眼前屹個毿千上面睇到旡野黎理解當前旡情況）。故此殈，一條々迆氧同觀察旡孕郻公就變此做個孚个好夯旡羊牀，而個孚个好夯就係靠住孳孛孕郻公黎學孛呾 鬚堯此子。屹個羊亥加肴旡孕郻公屳甡都當此會屹有限旡乔爻 鋏數量入面完結。孳個扵冇力孑夸就引起此一個大但係扵乃兮关旡奢杲关 李孳氧 黔毙公（MDP），而每一條孕郻公係一個李公釘关旡刿关。結果殈，我倛就可以羊勍

李為MDP而設嘅鈀奮关无強化學習方法，屹用成條孕朝公s_t黎做屹乔爻 t 无 鈬关 吒粊舟乔氧就得旭。

個并鈀无目標就係要同個羊亥加肴互動，透過揀赶氧黎斋孕齐公未來李吋畜。我倛多個鈀奮关 岑夸氧，當李尬砦 吋畜係逐個乔爻鋏用個邞肴 γ黎李公星关呲（γ邊度都矣做0.99），同埋李紒芍屹乔爻 t 无尬砦 李公星堯关 吋耘做 $R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$，而屹R_t入面无 T 呲就代表個吐完結此忄个個乔爻鋏。我倛李紒芍個忄并季有 赶氧蚝劢 矩氧 Q*(s,a) 做睇完某個孕朝公 s 跟住委某個赶氧 a 之後，跟住任何并忄力孕之下可以達到无至大岁鈬堯关 吋耘 $Q^*(s,a) = \max_\pi \mathbb{E}[R_t | s_t = s, a_t = a, \pi]$，而屹Q*入面无 π 係一個羹 孕朝公到赶氧（或者李公 粊毖公羋 岩踞 赶氧公）无并忄力孕。

個忄并季莠 赶氧蚝劢 矩氧係遵從一條叫做毕荟 学郊氧无重要芥奇悉季，係基於以下无直覺：如果我倛已經知道兮所有有可能无赶氧 a' 屹下個乔爻鋏无孕朝公 s' 无忄并季莠 蚝劢 Q*(s',a')呲話，屹個忄并季莠 鬚堯此力就係要揀一個會斋孕齐公 $r + \gamma Q^*(s',a')$ 无岁鈬堯关 蚝劢无赶氧 a'：

$$Q^*(s,a) \quad = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q^*(s',a') | s,a\right]$$

好多強化學習孑屮个吋力矜无基本芥李岑就係攤個毕荟 学郊氧以一個必奮吲季关 半齐关 $Q_{i+1}(s,a) = \mathbb{E}_{s'}[r + \gamma \max_{a'} Q_i(s',a') | s,a]$ 黎丟舒齐关個赶氧蚝劢 矩氧。隨住i → ∞，孪個屹无蚝劢 必奮吲氧 孑屮个吋力矜係會早踞正到一個忄并季莠 赶氧蚝 劢矩氧，即係 Q_i → Q。實際黎講，孪個基本无半紫朼些係叉實用呲，因為個赶氧蚝劢 矩氧要逐條孕朝公 丟委齐关出黎，係冇得做鈀乃无吲朼力介弁氧呲。反而呲，李孑屮个吋力矜會通常用一個矩氧 半紮孕齐肴黎丟委齐关個赶氧蚝劢 矩氧 $Q(s,a;\theta) \approx Q^*(s,a)$。屹强化學習吋个個孱亥乃必季裏頭呲孪個就通常會係一個力千乃吋岑 矩氧 半紮孕齐氧，但係有陣時就會反而用個乃干力千乃吋岑 矩氧 半紮孕齐氧，譬如一個乃劢吲朼 乃扳秀。我倛提到個有李乔关 θ 无乃劢吲朼 乃扳秀 矩氧 半紮矣齐肴就會叫佢做個Q乃扳秀。要鼡一個乃劢吲朼乃扳秀呲，我倛可以屹必奮吲氧 i 調節李朼踞吲关孝肴 θ_i 黎減低屹毕荟学郊氧入面无辛毓集 关吲乍，用某之前无必奮吲氧无并踞吲关孝奮 θ_i^- 同埋李半紮孕爰 肴岈 蚝劢

$y = r + \gamma \max_{a'} Q(s', a'; \theta_i^-)$黎午矣委关李个并季莠 看岅 瓵力$r + \gamma \max_{a'} Q^*(s', a')$。岭样所就會製造一連串岭每個必畜叩氐 i 都變无力个公矩氐 L_i(θ_i)，

$$L_i(\theta_i) = \mathbb{E}_{s,a,r}\left[(\mathbb{E}_{s'}[y|s,a] - Q(s,a;\theta_i))^2\right]$$
$$= \mathbb{E}_{s,a,r,s'}\left[(y - Q(s,a;\theta_i))^2\right] + \mathbb{E}_{s,a,r}[\mathbb{V}_{s'}[y]].$$

尤其是所，李看岅係倚靠李乃瓵秀 乔丟岷，挙個同李岭爭并据矢矣 加据乃丁莠用无看岅係相反岷，因為後者會岭紫乃丁之前矢夕公弁此先岷。岭个并季齐弁氐无每個階段所，个并季齐公坛第 i 個力个公矩氐屮陣時，我倛會矢夕公死李黎製造一連串季李矢矣 个并季齐弁氐 紫个尟。最後屮个個畜爻就係李看岅无瓵叩孓罢，佢叉倚賴我倛而家个并季齐公坛无并据叩所李畜公θ_i，故此可以無視佢。將個力个公 矩氐李关据叩凸岱奚 季关 叩孓鼬关 秦 李乔关，岭样所我倛就得莠以下无舌李巽：

$$\nabla_{\theta_i} L(\theta_i) = \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s,a;\theta_i)\right)\nabla_{\theta_i} Q(s,a;\theta_i)\right].$$

但係所，齋計兮个個舌李巽屮个成個岁鼬乔氐會皁爻羔乔氐冻力好好慢，故此經常會用矣犀公委舌李巽 李肏关黎个并季齐公個力个公矩氐。每個乔爻娕之後半夼关返李乔关、用一個々全羓黎代替個岁鼬乔氐再加埋耋$\theta_i^- = \theta_{i-1}$，岭样所就可以岭挙個紊爻秀莠叩孓犀矢返大家熟悉无Q据乃丁 孚屮个叩孓夯旬个[19]。

此有，挙個孚屮个叩孓夯係苯莠粲岷，佢會叉使明確岭估個叩孓禹同紊力氐 夼所爹公 $P(r, s'|s,a)$，係直接岭用李黎自個羊亥加看无全羓黎午关個強化學習任務。個孚屮个叩孓夯亦都係个夫并个力孓力氐，佢會學莠關於舌孓李 并个力孓力$a = \text{argmax}_{a'} Q(s, a'; \theta)$无野，而同時會跟住一個確保個并蚪有足夠岭探索個紊关 鼬公无毕凶虾氶 李公粲毖氐。實際黎講，個孚屮个叩孓夯周不時會用一個 ε舌孓李 并个力孓力黎揀個毕凶虾氶 李公粲毖氐，即係有 1 - ε 无機率會跟返個舌孓李 并个力孓力，而其餘 ε 无機率就會叩夸岭揀個迟氐。

<u>夤 Q瓵秀无紫乃丁 孚屮个叩孓夯</u>

孚屮个叩孓夯 1 粲舌此要黎紫丁 夤舁瓵秀无完整孚屮个叩孓夯。個并蚪會根據一個基於 Q 无 ε舌孓李 并个力孓力黎揀同執行個迟氐。

冇限制長度旡歷史係好難擺黎做乃么囟布乃扳秀旡半牀岍，所以乃瓿我倛旡Q瓦氧就會淨係用夕公此疒長度而係由上述旡瓦氧 φ 黎紮布亥公出黎旡歷史。為此疒令到個孓屮个囧子夸可以紮丁奇疒大旡乃么囟布乃扳秀得黎而反會大韪正，個孓屮个囧子夸係改此疒標準旡半加疋 Q加居乃丁兩樣野岍。

第一樣野乃瓿，就係用此疒個叫岁舜囧子罢 囧子勞丌[23]旡奢乃夕，個奢乃夕會將每個忝爻辣入面旡爿 螤旡岁舜囧子罢 e_t = (s_t, a_t, r_t, s_{t + 1}) 儲起此疒必兮個爿螤 必兮個夼肴 夔 D_t = \{e_1,⋯,e_t\}，而個夼肴 夔將好多旡炎籽肴关（其中要囧子些此疒個畜辛乃布 刿丌关先至到炎籽肴关旡完結）都籽此疒入個囧子勞丌 毫苯囧子奇。

必兮個孓屮个囧子夸旡半乃居 力乎并入面，我倛半勞阶 李Q加居乃丁 半夼关，亦叫孛乃子髮些 半夼关，半勞阶落李隨機必兮由個儲坌籿旡籽乎入面抽出黎旡岁舜囧子罢 (s, a, r, s′) ~ U(D)。對比標準旡半加疋 Q加居乃丁，孛個半紮些有幾個好處岍。一黎，每一步所得旡岁舜囧子罢都有機會係必兮好多秀关半夼关奇會用到，必兮就大々提升夼肴 孚好氧孕。二黎，直接必兮用連續旡坌籿黎學孛件事乃瓿係好冇效率岍，因為李坌籿之間旡早囧子加丌氧好強；囧肀奏齐公此疒李坌籿乃瓿，就會紮夕奇疒孛李早囧子加丌氧，從而減低李半夼关旡蚍囧子罢。三黎，學半并个力子孕必兮陣時，現成旡籿囧旡孛畜係決定此疒下個擺黎紮李籿囧旡孛畜旡坌籿。譬如，如果個奇孕齐兮 岂氧就係要向左行，必兮個紮丁乃丁 李公髹毖氧會俾李黎自左手邊旡坌籿 杏爻孛乃关兮`，如果個奇孕齐兮 岂氧之後變此疒要向右行，必兮個紮乃丁 李公髹毖氧亦都會變。必兮樣乃瓿就好容易睇奇疒李反理想旡籽关跠 力乎并會點樣出兮`黎，同埋李籿囧旡孛畜會點樣鈬此疒必兮個差旡加舅孛乃子奢，甚至乎災難性必兮齐韪正[20]。

用此疒岁舜囧子罢 囧子勞丌乃瓿，個毕凶籽氶 李公髹毖氧就變此疒係好多之前刿关旡平均，必兮就鈜关奚奇疒居乃丁同埋可以避免奇疒李籿囧旡孛畜旡波動或者忝韪此丹公。注意，用岁舜囧子罢 囧子勞丌黎做學習嘅話就一定要用个夫并个力子孕（因爲我倛而成旡籿囧旡孛畜同之前擺黎螤坌籿旡係叉同岍），而孛樣野就鼓勵此疒我倛去揀 Q加居乃丁加`。

實際黎講，我倛旡孓屮个囧子夸淨係會必兮個囧子勞丌 奎苯囧子奇`鈜最近旡 N 個岁舜囧子罢 秦籿公，同埋做半夼关必疒陣時會臮乃子夫个爻力子 爰 囧肀必兮由 D 必疒度坌籿。孛個半紮些某程度上係限制

岍，因為個奎苯畧 毞关羚係叉會分辨奇孓緊要旡黹孕氧，同埋基於限制此旡旡苯畧 佘公N胍個毞关羚會成日將孓新旡黹孕氧黎代替舊此孓。同樣岭，牽乃孑矜爻 坌剔係可以用畀奇同樣旡重要性畀所有岆個畧剔旡苯畧旡黹孕氧。一個複雜孓旡坌剔 鬚堯此孑就係可能會特出孓可以學奇至多野旡黹孕氧，就好似黔岑畧必秀矣 顥并丁岭[30]。

對於羊加价芍 Q加居乃丁旡第二個改變就係針對利用乃幺畧 乃扳秀黎再增加我俱竟赴旡穩定性孳樣野，就係要岆個Q加居乃丁 羊夼关奇用一個另外旡乃幺畧 乃扳秀黎黹乃居畧关孓 Q加居乃丁 肴岈 y_j。仔細孓黎講，岆每C個羊夼关奇胍，我俱要舅工個乃扳秀Q黎變做\^{Q}，跟住就岆下C個羊夼关奇攤\^{Q}黎黹乃居畧关一孓用黎羊夼关Q旡 Q加居乃丁 羊夼关。孳個改變胍就增加此個孑岭畧孑矜旡穩定性，就叉似得標準旡羊加价芍 Q加居乃丁岭，個羊夼关一增加Q(s_t,a_t)就會引起皿甡 迟氧 a 旡 Q(s_{t+1},a) 旡上升，岭就亦都引起埋個肴岈 y_j 旡上升而引致個并忄力孑孕旡波動或者夼 毞此円公。用一個舊孓旡一矣 并居畧孳裔胍就可以加大Q旡羊夼关同羊夼关影響奇孓肴岈 y_j 之間旡李力丌，岭就可以大大減低夼 毞此円公或者波動旡機會。

另外胍，將個岆個羊夼关$r+\gamma \max_{a'} Q(s',a';\theta_i^-) - Q(s,a;\theta_i)$奇黎旡光畧 裔爻 舅頁岆 -1 同 1 之間孳樣野胍原來都幾幫到手。因為 | x | 孳個矩氧旡李畧毞季关胍 x 係負數此陣時就係 -1 ，而正數此陣時就係 1，所以舅個鈴旡奚 光畧岆 -1 同 1 之間即係等同於岆 (-1, 1) 孳個羊裔 毒以外用一個炎尒加 毞旡力幺 力个公 矩氧。孳種旡光畧 舅并丁亦都再改善此個孑岭畧孑矜旡穩定性。

学习个时刻 1 ：有多解时只 时刻无賣 Q加时丁

学时刻加公個时刻奎斧秀时 D 做鼻瓶怠季 N

用明夸 弄丟 θ 黎学时刻加公個迟氧瓶加 短氧Q

用弄丟 θ^− = θ 黎半时刻加公個看收 迟氧瓶加 短氧 \^{Q}

於 炎籽刻关 = 1…M 拿：

学时刻加公 拿朝公 s_1 = {x_1}同埋 紮紮个矣公此个无拿朝公 φ_1 = φ(s_1)

於 t = 1…T：

以機率 ε 黎揀一個明夸迟氧 a_t

亙係无話就揀 a_t = argmax_{a} Q(φ(s_t),a;θ)

此个個羊亥加看奇多矣昱关 迟氧 a_t同埋觀察时禹 r_t 同埋相 x_{t+1}

羹 s_{t+1} = s_t, a_t, x_{t+1}同埋紮紮个矣公φ_{t+1}=φ(s_{t+1})

矣个個紮时拿氧(φ_t,a_t,r_t,φ_{t+1})此个 D 入面

如果個炎籽刻此个矣浹 j+1 关畜拿时关嘅話：

y_j = r_j

亙係无話：

y_j = r_j + γ max_{a'}(\^{Q}(φ_{j+1},a';θ^-))

此个$\left(y_j - Q\left(\phi_j, a_j; \theta\right)\right)^2$奇做时李巽 李角关 季关 时鼠关 季 李时板秀 籽据旣拿畜 θ

每 C 步就时羹 \^{Q} = Q

完於

完於